JOURNAL OF MULTIVARIATE ANALYSIS 40, 205–212 (1992)

Two LDF Characterizations of the Normal as a Spherical Distribution*

THEOPHILOS CACOULLOS

University of Athens

Communicated by the Editors

Two optimal characteristic properties of the normal distribution are shown: (a) Of all the SNM (spherical scale normal mixtures) the normal with the same Mahalanobis distances between Π_i: SNM(μ_i) and Π_j: SNM(μ_j), $i \neq j$, maximizes the probabilities of correct classification determined by a certain subclass of the LDF classification rules; (b) The class of LDF (linear discriminant function) rules is the admissible class for the discrimination problem with spherical population alternatives iff the spherical distribution is normal.

1. INTRODUCTION AND SUMMARY

Recently attempts have been made to relax and possibly dispense with the assumption of normality, thus allowing for more realistic, platykurtic and leptokurtic, non-normal alternatives in multivariate statistical analysis. Thus, the robustness of many classical multivariate procedures, especially, large-sample approximations, have been studied for the family of spherically (or elliptically) contoured or symmetric distributions, characterized simply by the fact that their density is constant on spheres (or ellipsoids). These spherically symmetric or, simply, spherical distributions (SD) may be also called spherically *isopycnic*, Cacoullos [3], and are sometimes referred to as *isotropic* or *radial*, as well. For some fundamental properties of SD, refer to Muirhead [8].

The k-population classification problem with k alternative normal populations $N(\mu_i, \Sigma)$, with known parameters μ_i and Σ is identified with the well-known linear discriminant functions (LDF).

$$L_{ij} = \alpha'_{ij} x = x' \Sigma^{-1} (\mu_i - \mu_j), \qquad i \neq j, \quad i, j = 1, ..., k; \tag{1.1}$$

Received July 30, 1990; accepted August 2, 1991.

AMS 1980 subject classifications: 62H30, 62H05.

Key words and phrases: Spherical distributions, linear discriminant functions, characterizations of normality, spherical normal mixtures, discriminatory power.

* Work sponsored in part by AFOSR under Grant AFSO-88-0030 to the Center for Multivariate Analysis, University of Pittsburgh.

0047-259X/92 $3.00

actually, there are only $m=\min(p, d)$ linearly independent LDFs; d $(1 \leqslant d \leqslant k-1)$ denotes the dimensionality of the linear space E_d spanned by the k points $\mu_1, ..., \mu_k$ in the p-space of x, the observation to be classified into one of the k populations.

In the case of maximal dimensionality $(d=k-1,\ p \geqslant k-1)$, considered here, the simplicity and optimality (Bayes and admissibility) of the k classification regions (see, e.g., Anderson [1, Chap. 6])

$$R_i\colon L_{ij} \leqslant c_{ij}, \qquad j \neq i \quad (i=1, ..., k) \tag{1.2}$$

(x is classified into the ith population iff $x \in R_i$) raises the question about the feasibility and performance (discriminatory power) of these LDF regions R_i when the assumption of normality is replaced by spherical symmetry (since Σ is known, we can take $\Sigma = I_p$ and elliptical symmetry reduces to spherical symmetry). In particular, the feasibility and evaluation of the minimum-distance rule was examined by Koutras [5, 6].

We show that certain optimality properties of the LDF regions R_i characterize normality in the SD family.

In the first place, the results concern the important subfamily of SDs, the so-called spherical normal (scale) mixtures (SNM), defined by (2.6). More specifically, it is shown that, within the SNM family the normal with the same Mahalanobis distance between two alternative populations maximizes the probabilities of correct classification (PCC) for any classification rule (R_1, R_2) with minimum PCC $\frac{1}{2}$ (Theorem 2.1). Similar optimality properties of normality are discussed for $k>2$ populations, but the results are not so neat, due to the complexity of the region of concavity of the multivariate normal CDF.

Second, the question of admissibility of the regions R_i for the location-discrimination problem within the SD class is considered, and it is shown (Theorem 2.2) that the class of R_i rules coincides with the class of (Bayes) admissible rules (under a simple 0–1 loss) iff the alternative SDs are normal.

2. Motivation and Main Results

Let Π_i: SD(μ_i, g) denote a p-dimensional spherical distribution with density

$$f_i(x) = cg(|x-\mu_i|^2), \qquad i=1, ..., k, \tag{2.1}$$

with given location parameters (means, if they exist) $\mu_1, ..., \mu_k$; $g \geqslant 0$ is such that

$$\int_0^\infty t^{(p/2)-1} g(t)\, dt < \infty$$

and c a normalizing constant depending on p and g. Given an observation $x=(x_1, ..., x_p)$, we wish to classify it into one of the k alternative SD populations.

The discriminatory power of the minimum distance (MD) classification rule $R_0=(R_1^0, ..., R_k^0)$, according to which an observation x is classified into population Π_i iff $x \in R_i^0$, where

$$R_i^0=\{x \in R^p\colon Q_i(x) \leqslant Q_j(x), j \neq i),$$

and $\qquad i=1, ..., k, \qquad (2.2)$

$$Q_i(x)=(x-\mu_i)'(x-\mu_i)=|x-\mu_i|^2,$$

has been evaluated by Koutras [6] in the simple case $k=2$. Independently of the admissibility (see (2.2) of [3]), he showed that the maximum probability of correct classification (PCC) is an increasing function $P(\delta)$ of $\delta=|\mu_1-\mu_2|$, for any given pair $\Pi_1\colon \mathrm{SD}(\mu_1, g)$ and $\Pi_2\colon \mathrm{SD}(\mu_2, g)$ of the same family g. Given a δ and two different g's, e.g., the normal with

$$g(t)=g_0(t)=e^{-rt}, \qquad r>0, \tag{2.3}$$

and the p-variate Student t_n with n degrees of freedom, i.e.,

$$g(t)=g_{1,n}(t)=\left(1+\frac{t}{n}\right)^{-(n+p/2)}, \tag{2.4}$$

we observe that, e.g., if $n=1$ (Cauchy) and g_0 the standard normal $(r=\frac{1}{2})$, the Cauchy appears to give a higher PCC than the normal (see Koutras [6] for specific values of $P(\delta)$), that is,

$$P(\delta \mid g_0)=\Phi\left(\frac{\delta}{2}\right)<P(\delta \mid g_{1,1}), \qquad \delta>0;$$

this is not true, of course, if $n>2$, when the dispersion matrix of t_n exists and $P(\delta \mid g_{n,1})$ tends to $\Phi(\delta/2)$ as $n \to \infty$. Similar behavior is exhibited when we compare $P(\delta \mid g_0)$ with $P(\delta \mid g_2)$, where g_2 denotes the generalized Laplace or Bessel SD with

$$g(t)=g_{\alpha,s}(t)=\left(\frac{\sqrt{t}}{s}\right)^{\alpha} K_\alpha\left(\frac{\sqrt{t}}{s}\right), \qquad \alpha>-\frac{p}{2}; \tag{2.5}$$

in particular, when $\alpha=1-p/2$, $P(\delta)$ simplifies to (see (4.8) in Koutras [6])

$$P(\delta)=1-\tfrac{1}{2}\exp[-\delta/2s], \tag{2.5*}$$

which for small enough s again exceeds $\Phi(\delta/2)$.

It turns out that the observed differences in $P(\delta)$ are not only due to the difference of g's, but also to different Mahalanobis distances Δ; indeed, for g_0, $\Delta=\delta$, whereas under (2.5), with $\alpha=1-p/2$, the dispersion matrix is $s^2 I_p$ and the corresponding $\Delta=\delta/s$, which agrees with the implications of (2.5)*.

The preceding observations and the fact that (2.4) and (2.5) belong to the special class of SD, the so-called spherical (scale) normal mixtures (SNM), defined by (see (2.3))

$$g(t)=\int_0^\infty e^{-rt}\,dF(r), \tag{2.6}$$

with F a distribution function on $[0, \infty)$, motivated, in the first place, the result, Cacoullos [3], that within the SNM family the maximum PCC (minimax misclassification probability) for given Mahalanobis distance Δ is maximized by the normal.

Actually a stronger result holds for two populations:

THEOREM 2.1. *Let $\Pi_i(F)$: $\mathrm{SNM}(\mu_i, g)$, $i=1, 2$, i.e., g as in* (2.6). *Then, of all $g \in \mathrm{SNM}$ the normal with the same Mahalanobis distance Δ between Π_1 and Π_2 maximizes the probabilities of correct classification of any rule (R_1, R_2) such that*

$$R_1=R_1(\lambda)=\{x\colon \delta_{12}(x)\leqslant \lambda\Delta^2\}, \qquad -1<\lambda<1, \tag{2.7}$$

where $\delta_{ij}(x)=Q_i(x)-Q_j(x)$ and $Q_i(x)$ was defined in (2.2).

Proof. If $g(t)=g_0$ and $r(t)=e^{-rt}$, r fixed, then, $\delta_{12}(X)=(2X-[\mu_1+\mu_2])'(\mu_2-\mu_1)$ is $N(\delta_{12}(EX), 4\Delta^2(r))$, where the Mahalanobis distance $\Delta=\Delta(r)$ between Π_1: $N(\mu_1, (2r)^{-1} I_p)$ and Π_2: $N(\mu_2, (2r)^{-1} I_p)$ is $\Delta(r)=\sqrt{2}\, r\delta$, so that the corresponding probabilities of correct classification, (CC), are

$$\begin{aligned} P_1=P[\mathrm{CC}\mid \Pi_1]&=P[\delta_{12}(X)\leqslant \lambda\Delta^2(r)\mid \mu=\mu_1]=\Phi\left(\frac{1+\lambda}{2}\Delta(r)\right),\\ P_2=P[\mathrm{CC}\mid \Pi_2]&=P[\delta_{12}(X)> \lambda\Delta^2(r)\mid \mu=\mu_2]=\Phi\left(\frac{1-\lambda}{2}\Delta(r)\right). \end{aligned} \tag{2.8}$$

Hence the corresponding PCC under $\Pi_1(F)$ and $\Pi_2(F)$, by (2.6), are

$$P[\mathrm{CC}\mid \Pi_1(F)]=E_F\left[\Phi\left(\frac{1+\lambda}{2}\Delta(r)\right)\right],$$

$$P[\mathrm{CC}\mid \Pi_2(F)]=E_F\left[\Phi\left(\frac{1-\lambda}{2}\Delta(r)\right)\right],$$

E_F denoting expectation under the mixing distribution F of r. The assertion now follows by the Jensen inequality and the concavity of $\Phi(u)$ iff $u>0$ ($\Delta = E_F[\Delta(r)]$).

Taking $\lambda = 0$ in (2.7) yields the result for the MD rule, i.e., the minimax classification procedure, with minimax probability of error $1-\Phi(\Delta/2)$.

COROLLARY 2.1. *Of all* $g \in \mathrm{SNM}$ *with the same Mahalanobis distance* $\Delta > 0$ *between* Π_1: $\Sigma\mathrm{NM}(\mu_1, g)$ *and* Π_2: $\mathrm{SNM}(\mu_2, g)$, *the normal minimizes the minimax misclassification probability.*

Another way of stating Theorem 2.1 is

COROLLARY 2.2. *Let* Π_i *and* R_i *be defined as in Theorem* 2.1. *Then, within the* SNM *family the normal with the same Mahalanobis distances* $\Delta > 0$ *between* Π_1 *and* Π_2, *maximizes both probabilities* P_1 *and* P_2 *of correct classification, provided each of them is at least* $\frac{1}{2}$.

Let us now look at the situation with $k>2$ SNM alternatives. Clearly, if for $k=2$ only a certain subset (see (2.7)) of the LDF regions, R_i, defined by (1.2) (or (2.9) below), give higher PCC under normality than any other SNM with the same Mahalanobis distance, it should not now be expected to make any omnibus statement about the corresponding PCC. Consider, e.g., the simpler case of $k=3$; it is easily seen that, the joint distribution of the classification statistics $\delta_{ij}(X) = Q_i(X) - Q_j(X)$ (cf. Cacoullos [2]) is bivariate normal, with the cosines of the angles of the triangle (μ_1, μ_2, μ_3), i.e.,

$$\rho_i = \mathrm{Corr}(\delta_{ij}(X), \delta_{ik}(X)) = \frac{(\mu_i - \mu_j)'\,(\mu_i - \mu_k)}{|\mu_i - \mu_j|\,|\mu_i - \mu_k|} \qquad (i \neq j \neq k).$$

as correlations. It is thus obvious, in view of the simple case of Theorem 2.1, that a PCC superiority under normality will depend on the region of concavity of a bivariate normal CDF $F(x, y; \rho)$, say, of two jointly normal standard variables with correlation ρ. Thus, for fixed r and distances $\Delta_{ij} = \Delta_{ij}(r)$, between Π_i: $\mathrm{SNM}(\mu_i, g)$ and Π_j: $\mathrm{SNM}(\mu_j, g)$, $i \neq j$ (g as in (2.6)), we find for the regions (cf. (2.7) and (2.8))

$$R_i = \{x \in R^p \colon \delta_{ij}(X) \leqslant \lambda_{ij}\Delta_{ij},\ j \neq i\}, \qquad i = 1, 2, 3, \tag{2.9}$$

that, e.g.,

$$P_1 \equiv P[\mathrm{CC} \mid \Pi_1] = F\left(\frac{1+\lambda_{12}}{2}\Delta_{12}, \frac{1+\lambda_{13}}{2}\Delta_{13}; \rho_1\right). \tag{2.10}$$

Partial answers, for the desired comparisons of the P_i under normality and general SNM, are provided by the following sufficient conditions for the concavity of $F(x, y; \rho)$.

LEMMA 2.1. *Let $F(x, y; \rho)$ denote the* CDF *of two jointly normal variables X and Y with means* 0, *variances* 1, *and correlation ρ. Then $F(x, y; \rho)$ is concave for all (x, y) such that $x > \gamma_0$, $y - \rho x > \gamma_0$ or $x - \rho y > \gamma_0$, $y > \gamma_0$, where $\gamma_0 = 0.506$ is the solution of $x\Phi(x) = \phi(x)$.*

This is a simple corollary of the orthogonality of X and $Y - \rho X$ and the following result, perhaps of some interest in itself.

PROPOSITION 2.1. *The* CDF $\Phi^*(x) = \prod_{i=1}^{p} \Phi(x_i)$ *of* $X \sim N(0, I_p)$ *is concave if $x_i > \gamma_0$ for each $i = 1, ..., p$.*

The proof consists in verifying that the Wronskian determinant of Φ^* is positive, whereas the diagonal elements are negative, when all the $x_i > \gamma_0$.

The result for Φ^* can be used to yield sufficient conditions for the concavity of the CDF $F(x; \rho_{ij}, i \neq j)$ of $X \sim N(0, R)$, $R = (\rho_{ij})$, $\rho_{ij} = \text{Corr}(X_i, X_j)$. Thus, as for $F(x, y; \rho)$ above, one has to orthonormalize $X_1, ..., X_p$, and then seek conditions on λ_{ij} and Δ_{ij} so that the CDF corresponding to R_i (cf. (2.9)), i.e.,

$$F\left(\frac{1 + \lambda_{ij}}{2} \Delta_{ij}, j \neq i; \rho_{ij}, j \neq i\right) = P[\text{CC} \mid \Pi_i], \qquad i = 1, ..., k,$$

is concave.

As an application of Lemma 2.1, let us return to the situation of (2.10) and examine the MD rule, obtained by setting $\lambda_{12} = \lambda_{13} = 0$; its PCC under Π_1 is given by $F(x, y; \rho_1)$ with $x = \Delta_{12}/2$, $y = \Delta_{13}/2$, $\rho_1 = \cos \theta_1$ (θ_1 is the angle at μ_1 of the triangle (μ_1, μ_2, μ_3)). Using Lemma 2.1 and taking, without any loss of generality, $\Delta_{12} \leqslant \Delta_{13} \leqslant \Delta_{23}$, we arrive at

COROLLARY 2.3. *Under the conditions of* (2.7), *the normal,* SD *of all* SNM *with the same Mahalanobis distances Δ_{12}, Δ_{13}, Δ_{23}, maximizes the $P[\text{CC} \mid \Pi_1]$ determined by the* MD-*rule* $R_0 = (R_1^0, R_2^0, R_3^0)$, *provided that*

(a) $\Delta_{12} > 2\gamma_0$, $\Delta_{13} > 2\gamma_0 \sqrt{2}$ *if* $\cos \theta_1 > 0$ *or*

(b) $\Delta_{12} > 2\gamma_0$ *if* $\cos \theta_1 < 0$.

Hence, we conclude the following interesting

COROLLARY 2.4. *The normal of all the* SNM *with the same Mahalanobis distances $\Delta_{12} \leqslant \Delta_{13} \leqslant \Delta_{23}$ maximizes all three* PCC ($k = 3$) *determined by the* MD *rule, provided that*:

(a) $\Delta_{12} > 2\gamma_0$, $\Delta_{13} > 2\gamma_0 \sqrt{2}$, $\Delta_{23} > 2\gamma_0 \sqrt{2}$ *if the triangle (μ_1, μ_2, μ_3) is acute-angled*

(b) $\Delta_{12} > 2\gamma_0$, $\Delta_{23} > 2\gamma_0 \sqrt{2}$ *if* $\cos \theta_1 < 0$ (*i.e., the triangle (μ_1, μ_2, μ_3) is obtuse-angled*).

Finally, it may be of interest to examine the conditions under which the above optimality property of the normal holds for the minimax LDF classification procedure, which is well known to equalize the PCCs. In general, judging from the preceding results for $k=2$ and $k=3$, one might say that, roughly speaking, if the Mahalanobis distances are large enough so that the corresponding PCCs P_i are large and about the same (cf. minimax and MD rule), then the normal will have the above optimality property in the SNM family. We cannot, however, go into further details in this note.

LDF Discrimination Is Best Only under Normality

It is well known (see Anderson [1]), that if the alternative populations Π_i are normal, then the class of Bayes admissible classification regions R_i are of the form (1.2) or (2.9). They are also the admissible invariant (under the affine group) procedures for the composite-alternatives problem of selecting the nearest Π_i: $N(\mu_i, \Sigma)$ to an unknown Π_0: $N(\mu, \Sigma)$, as shown by Cacoullos [2].

The question raised here is whether the admissibility of R_i implies the normality of Π_i: $\mathrm{SID}(\mu_i, g)$, $i=1, ..., k$. The answer is affirmative, as shown in

THEOREM 2.2. *Let Π_i: $\mathrm{SD}(\mu_i, g)$, $i=1, 2$, and suppose the admissible regions R_1, R_2, are defined by* (1.2) *or* (2.7) *with $-\infty < \lambda < \infty$. Moreover, assume that g, in addition to being continuous, is also differentiable* (*almost everywhere*). *Then g is normal.*

Proof. An admissible region R_1 is a likelihood ratio region, that is, satisfies

$$R_1 = \left\{ x \in T^p : \frac{f_1(x)}{f_2(x)} = \frac{g(Q_1(x))}{g(Q_2(x))} > c \right\}, \qquad \text{for some} \quad c > 0, \tag{2.10}$$

so that, letting $h(x) = \ln g(x)$, (2.10) is satisfied iff

$$h(Q_1(x)) - h(Q_2(x)) = \Psi(Q_1(x) - Q_2(x)),$$

where (cf. (2.7) Ψ must be a decreasing function of $Q_1(x) - Q_2(x) = \delta_{12}(x)$). Letting $Q_2(x) \to Q_1(x)$, from the continuity of h we obtain $\Psi(0) = 0$, whereas from the differentiability of h (by hypothesis on g), we see that, for each t, the derivative $h'(t)$ is equal to the limit of $\Psi(t)/t$ as $t \downarrow 0$, i.e., $\Psi'(0)$, so that $h'(t)$ is a constant. Hence the assertion follows.

References

[1] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed., Wiley, New York.

[2] Cacoullos, T. (1965). Comparing Mahalanobis distances. I. Comparing distances between k known normal populations and another unknown. *Sankhyā A* **27** 1–22.

[3] Cacoullos, T. (1988). *On the Optimality of the Linear Discriminant Function for Spherically Isopycnic Distributions*, TR No. 88-06, Center for Multivariate Analysis, University of Pittsburgh.

[4] Cacoullos, T., and Koutras, M. (1985). Minimum-distance discrimination for spherical distributions. In *Statistical Theory and Data Analysis* (K. Matusita, Ed.), pp. 91–102, Elsevier, New York.

[5] Koutras, M. J. (1983). *Contributions to the Theory of Spherical Distributions and Related Classification Problems*. Doctoral thesis, University of Athens. [Greek]

[6] Koutras, M. (1987). On the performance of the linear discriminant function for spherical distributions. *J. Multivariate Anal.* **22** 1–12.

[7] Muirhead, R. (1984). *Aspects of Multivariate Statistical Theory*, Wiley, New York.